Is tracking all that it takes?

Exploring the validity of news media exposure measurements created with metered data

Oriol J. Bosch | Department of Methodology, LSE & RECSM - UPF

Melanie Revilla | RECSM - UPF

o.bosch-jover@lse.ac.uk

orioljbosch

https://orioljbosch.com/

Acknowledgements: I would like to thank Patrick Sturgis and Jouni Kuha

Funding: This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 849165; PI: Melanie Revilla); the Spanish Ministry of Science and Innovation under the "R+D+i projects" programme (grant number PID2019-106867RB-I00 /AEI/10.13039/501100011033 (2020-2024), PI: Mariano Torcal); and the BBVA foundation under their grant scheme to scientific research teams in economy and digital society, 2019 (PI: Mariano Torcal).

The rise of metered data to understand online media exposure

- **Two parallel trends:**
  1. Increasing importance of understanding what kind of media people are exposed to;
  2. Concerns about the data quality of self-reported exposure to media

The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure

Get access

Markus Prior

Public Opinion Quarterly, Volume 73, Issue 1, Spring 2009, Pages 130–143, https://doi.org/10.1093/poq/nfp002

Published: 18 March 2009

Cite Permissions Share

Abstract

Many studies of media effects use self-reported news exposure as their key independent variable without establishing its validity. Motivated by anecdotal evidence that people's reports of their own media use can differ considerably from independent assessments, this study examines systematically the accuracy of survey-based self-reports of news exposure. I compare survey estimates to Nielsen estimates, which do not rely on self-reports. Results show severe overreporting of news exposure. Survey estimates of network news exposure follow trends in Nielsen ratings relatively well, but exaggerate

UK Parliament

Business | MPs, Lords & offices | About | Get invo

UK Parliament

> Democracy under threat from 'pandemic of misinformation' online – Lords Democracy and Digital Technologies Co

Democracy under threat from 'pandemic of misinformation' online – Lords Democracy and Digital Technologies Committee

The UK Government should act immediately to deal with a 'pandemic of misinformation' that poses an existential threat to our democracy and way of life. The stark warning comes in a report published today by the Lords Committee on Democracy and Digital Technologies.

The report says the Government must take action 'without delay' to ensure tech giants are held responsible for the harm done to individuals, wider society and our democratic processes through misinformation widely spread on their platforms.

The Committee says online platforms are not 'inherently ungovernable' but power has been ceded to a "few unelected and unaccountable digital corporations" including Facebook and Google, and politicians must act now to hold those corporations to account when they are shown to negatively influence public debate and undermine democracy.

The Committee sets out a package of reforms which, if implemented, could help restore public trust and ensure democracy does not 'decline into irrelevance'.

The rise of metered data to understand online media exposure

- **Two parallel trends:**
  1. Increasing importance of understanding what kind of media people are exposed to;
  2. Concerns about the data quality of self-reported exposure to media

- Alternative: directly observe what people do online using digital tracking solutions, or *meters.*
  - **Group of tracking technologies**
  - **Installed on participants devices**.
  - **Collect traces** left by participants when **interacting with their devices online: e.g. URLs or apps visited**

- We call the resulting data: **metered data** → **+60 papers** published using metered data since 2016

Measuring online media exposure with metered data

Concept of interest → **Measurement**

Measuring online media exposure with metered data

Concept of interest → **Measurement**

- Measurements: **pieces of information** from the participants' tracked online behaviour that are **combined**, and sometimes **transformed**, to compute **a specific variable.**

Measuring online media exposure with metered data

Concept of interest → **Measurement**

- Measurements: **pieces of information** from the participants' tracked online behaviour that are **combined**, and sometimes **transformed**, to compute **a specific variable.**

↳ The time stamps of all visited URLs defined as news media articles

Measuring online media exposure with metered data

Concept of interest $\xrightarrow{\textbf{Validity}}$ **Measurement**

Most research seems to expect this relationship to be perfect, but there is no evidence

- Measurements: **pieces of information** from the participants' tracked online behaviour that are **combined**, and sometimes **transformed**, to compute **a specific variable.**

→ The time stamps of all visited URLs defined as news media articles

Many unclear questions to answer

webdataopp

Online news media exposure

1. **Define the list of URLs that can be defined as "online news media"**
   a) Select a list of online news media domains → no complete one, which one to choose?
   b) Select which domains to use within those lists → all? The most visited? How many?
   c) Is all the information from the domain relevant, or only some specific URLs should be considered?
2. **Define what is considered as being "exposed"**
   a) Should all visits to an URL/App be considered? Only those complying with a specific rule?
   b) Should visits be counted? Or the time of those visits?
   c) Should information from all devices be used? Or only from specific devices?
3. **Define the time frame used to compute the variables**
   a) How many days of tracking should be used?
   b) Should information be from before the survey, from after the survey, or from both before and after the survey (in case a survey is used).

This study

Research questions

- Does the convergent validity of online news media exposure measured with metered data fluctuate across design choices? (**RQ1.1**)?
- Does the predictive validity of online news media exposure measured with meted data fluctuate across design choices? (**RQ1.2**)
- What design choices have a higher impact on predictive validity? (**RQ2.1**)
- To what extent do different design choices affect the predictive validity of metered data measures? (**RQ2.2**)

Data

TRI-POL project - Overview

- Three wave survey combined with metered data at the individual level
- $\boxed{\text{Spain, Portugal, Italy}}$ + Argentina and Chile
- Netquest metered panels – Cross-quotas about gender, age, education and region
- Sample size: 993 (Spain), 842 (Italy), 818 (Portugal)
- Fieldwork: September 21 – April 22

Design choices identified

Online news media exposure

Characteristics	Our choices
List	Own, Tranco, Alexa, Cisco, Majestic
Top	10, 20, 50, 100, 200, All
Information	All domain level, subdomains defined as political
Exposure	1 second, 30 seconds, 120 seconds
Level	Visits, Time
Devices	Mobile & PC, PC only, Mobile only
Days of tracking	2, 5, 10, 15, 31
Survey period	Before, After, Before and After

3,573 potential combinations

- Which ones should be preferred?
- Which ones should be avoided?
- Does it even matter?

Assessing whether validity fluctuates across design choices

First, we study convergent validity across the three countries (RQ1.1)

- "Convergent validity describes the fit between independent measures of the same underlying concept" (Prior, 2013).

Essentially, if all variables were measuring the same concept, they should highly correlate with each other

Assessing whether validity fluctuates across design choices

First, we study convergent validity across the three countries (RQ1.1)

- "Convergent validity describes the fit between independent measures of the same underlying concept" (Prior, 2013).

  → Essentially, if all variables were measuring the same concept, they should highly correlate with each other

- We computed one correlation for each potential pair of variables → 6,349,266 unique correlations

Assessing whether validity fluctuates across design choices

webdataopp

First, we study convergent validity across the three countries (RQ1.1)

- "Convergent validity describes the fit between independent measures of the same underlying concept" (Prior, 2013).

→ Essentially, if all variables were measuring the same concept, they should highly correlate with each other

- We computed one correlation for each potential pair of variables → 6,349,266 unique correlations

↓

RQ1.1: To what extent do these correlations fluctuate?

Assessing whether validity fluctuates across design choices

web data opp

Second, we study predictive validity across the three countries (RQ1.2)

- "Predictive validity refers to the degree to which scores on a test or assessment are related to performance on a criterion or gold standard assessment" (Frey, 2018).

Measures closer to the theorised *true* relationship should be preferred. In practice, since the *true* value is unknown, people assume that higher is better.

Assessing whether validity fluctuates across design choices

web data opp

Second, we study predictive validity across the three countries (RQ1.2)

- "Predictive validity refers to the degree to which scores on a test or assessment are related to performance on a criterion or gold standard assessment" (Frey, 2018).

Measures closer to the theorised *true* relationship should be preferred. In practice, since the *true* value is unknown, people assume that higher is better.

- Political knowledge has been used as the most common gold standard when assessing the predictive validity of news media exposure.

Smith B, Clifford S, Jerit J. **TRENDS: How Internet Search Undermines the Validity of Political Knowledge Measures**. *Political Research Quarterly*. 2020;73(1):141-155. doi:10.1177/1065912919882101
Dilliplane, S., Goldman, S. K., & Mutz, D. C. (2013**). Televised exposure to politics: New measures for a fragmented media environment**. *American Journal of Political Science*, *57*(1), 236-248.
Prior, M. (2009). **Improving media effects research through better measurement of news exposure.** *The Journal of Politics*, *71*(3), 893-908.

Assessing whether validity fluctuates across design choices

Second, we study predictive validity across the three countries (RQ1.2)

- "Predictive validity refers … or assessment are related to performance on a criteri… 2018).

…nship should be preferred. In practice, since …at higher is better.

- Political knowledge has b… ndard when assessing the predictive validity of new…

Political knowledge was measured by asking **4 knowledge questions** about politics.

The questions covered **basic knowledge** about the **political system**, and knowledge about the **current cabinets** in each country.

The final variable is a **sum of all correct answers**, hence, it ranges from **0 to 4**.

Smith B, Clifford S, Jerit J. **TRENDS: How Internet Search Undermines the Validity of Political Knowledge Measures**. *Political Research Quarterly*. 2020;73(1):141-155. doi:10.1177/1065912919882101
Dilliplane, S., Goldman, S. K., & Mutz, D. C. (2013). **Televised exposure to politics: New measures for a fragmented media environment**. *American Journal of Political Science*, *57*(1), 236-248.
Prior, M. (2009). **Improving media effects research through better measurement of news exposure.** *The Journal of Politics*, *71*(3), 893-908.

Assessing whether validity fluctuates across design choices

webdataopp

Second, we study predictive validity across the three countries (RQ1.2)

- "Predictive validity refers to the degree to which scores on a test or assessment are related to performance on a criterion or gold standard assessment" (Frey, 2018).

  → Measures closer to the theorised *true* relationship should be preferred. In practice, since the *true* value is unknown, people assume that higher is better.

- Political knowledge has been used as the most common gold standard when assessing the predictive validity of news media exposure.

- For each variable, we ran a regression model with political knowledge as the dependant variable, and several common control variables.

  → **3,573** unique coefficients

Assessing whether validity fluctuates across design choices

webdataopp

Second, we study predictive validity across the three countries (RQ1.2)

- "Predictive validity refers to the degree to which scores on a test or assessment are related to performance on a criterion or gold standard assessment" (Frey, 2018).

  → Measures closer to the theorised *true* relationship should be preferred. In practice, since the *true* value is unknown, people assume that higher is better.

- Political knowledge has been used as the most common gold standard when assessing the predictive validity of news media exposure.

- For each variable, we ran a regression model with political knowledge as the dependant variable, and several common control variables.

  → **3,573** unique coefficients → **RQ1.2:** To what extent do these coefficients fluctuate?

The impact of each design choice on predictive validity (RQ3)

- The variables were used as the observations, their associated **regression coefficients** as the dependant variable, and the characteristics of the variable as the predictors

→ Similar approach as for the *Survey Quality Predictor* (SQP)

Coefficients_dataset_PT

Filter

	variable	Coefficient	List	TOP	Level	Time_visit	Time_frame	PRE_POST	DOMAIN_SUBDOMAIN	Device
1	avgALL_T_News_100A	0.14578984	Alexa	100	Time	1	31	PRE_AND_POST	Domain	All devices
2	avgALL_T_News_100C	0.14720057	Cisco	100	Time	1	31	PRE_AND_POST	Domain	All devices
3	avgALL_T_News_100M	0.14772164	Majestic	100	Time	1	31	PRE_AND_POST	Domain	All devices
4	avgALL_T_News_100T	0.14542314	Tranco	100	Time	1	31	PRE_AND_POST	Domain	All devices
5	avgALL_T_News_10A	0.11781648	Alexa	10.	Time	1	31	PRE_AND_POST	Domain	All devices
6	avgALL_T_News_10C	0.12287777	Cisco	10.	Time	1	31	PRE_AND_POST	Domain	All devices
7	avgALL_T_News_10M	0.12597311	Majestic	10.	Time	1	31	PRE_AND_POST	Domain	All devices
8	avgALL_T_News_10T	0.12597311	Tranco	10.	Time	1	31	PRE_AND_POST	Domain	All devices
9	avgALL_T_News_200A	0.14578984	Alexa	200	Time	1	31	PRE_AND_POST	Domain	All devices
10	avgALL_T_News_200C	0.14720057	Cisco	200	Time	1	31	PRE_AND_POST	Domain	All devices
11	avgALL_T_News_200M	0.14772164	Majestic	200	Time	1	31	PRE_AND_POST	Domain	All devices
12	avgALL_T_News_200T	0.14542314	Tranco	200	Time	1	31	PRE_AND_POST	Domain	All devices
13	avgALL_T_News_20A	0.14319744	Alexa	20	Time	1	31	PRE_AND_POST	Domain	All devices
14	avgALL_T_News_20C	0.14519358	Cisco	20	Time	1	31	PRE_AND_POST	Domain	All devices
15	avgALL_T_News_20M	0.14372789	Majestic	20	Time	1	31	PRE_AND_POST	Domain	All devices
16	avgALL_T_News_20T	0.14335666	Tranco	20	Time	1	31	PRE_AND_POST	Domain	All devices
17	avgALL_T_News_50A	0.14578984	Alexa	50	Time	1	31	PRE_AND_POST	Domain	All devices
18	avgALL_T_News_50C	0.14720057	Cisco	50	Time	1	31	PRE_AND_POST	Domain	All devices
19	avgALL_T_News_50M	0.14772164	Majestic	50	Time	1	31	PRE_AND_POST	Domain	All devices
20	avgALL_T_News_50T	0.14778072	Tranco	50	Time	1	31	PRE_AND_POST	Domain	All devices
21	avgALL_T_News_ALL	0.15279798	ALL	222	Time	1	31	PRE_AND_POST	Domain	All devices

http://sqp.upf.edu/

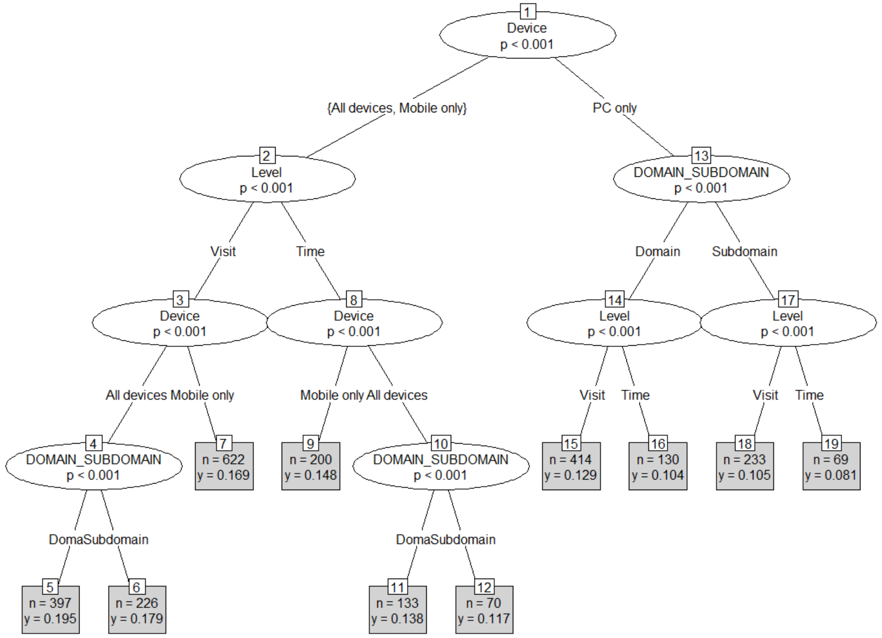
THIS STUDY

The impact of each design choice (RQ2)

- To predict the impact of each design choice, we used random forests of regression trees* (*randomForest* R package).

- Why?
  - Non-linearity
  - Correlated features
  - Feature importance
  - Ensemble learning

- We extract the following information:
  - The variable importance: % increase of MSE (**RQ2.1**)
  - And the marginal effect of each choice (**RQ2.2**)



* *Ntree*: 500 | *Mtry*: 6 | *Node size*: 3 | *Sample fraction*: 80%

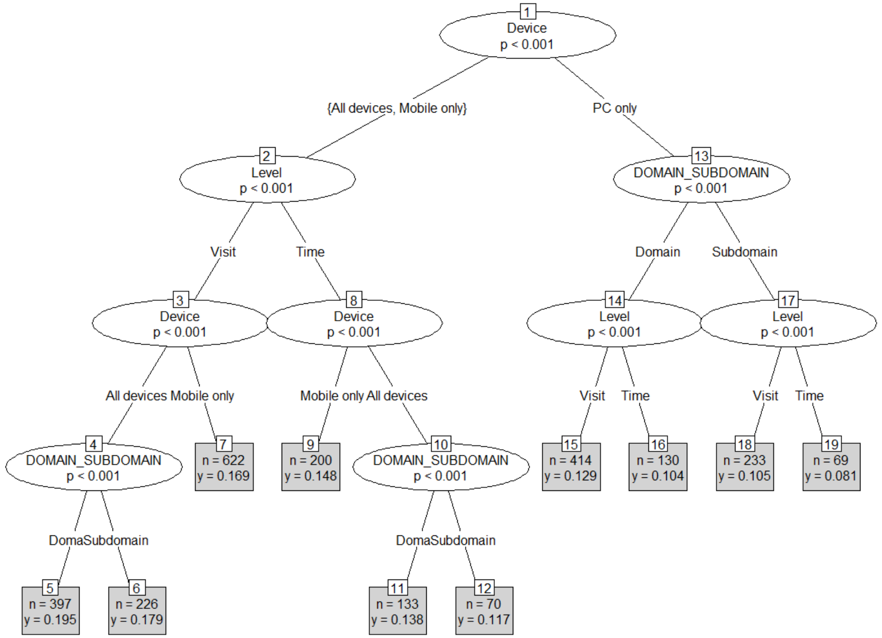
THIS STUDY

The impact of each design choice (RQ2)

- To predict the impact of each design choice, we used random forests of regression trees* (*randomForest* R package).

- Why?
  - Non-linearity
  - Correlated features
  - Feature importance
  - Ensemble learning

- We extract the following information:
  - The variable importance: % increase of MSE (**RQ2.1**)
  - And the marginal effect of each choice (**RQ2.2**)



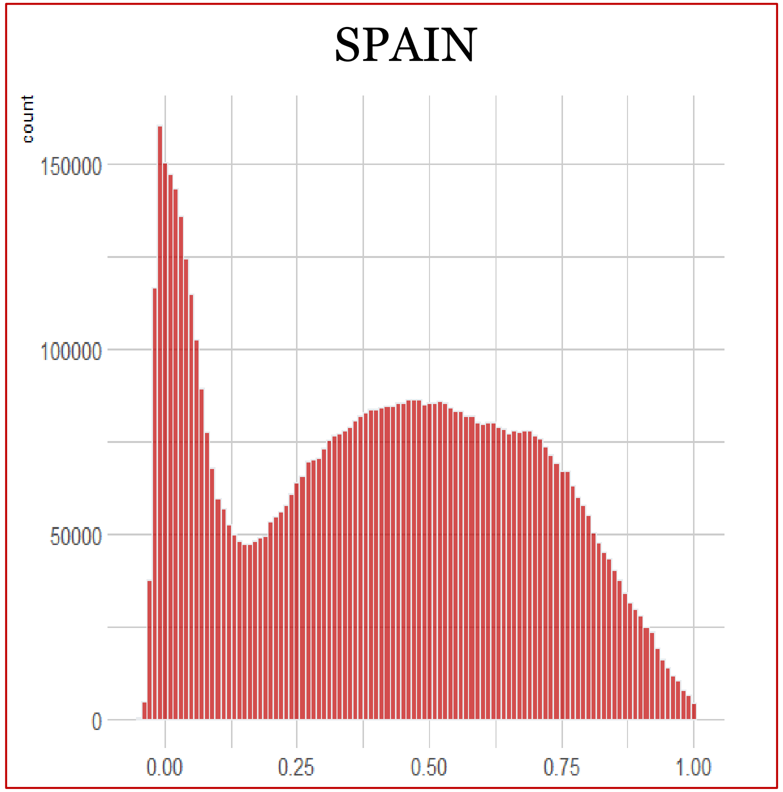
* *Ntree*: 500 | *Mtry*: 6 | *Node size*: 3 | *Sample fraction*: 80%

Does the validity of online news media exposure measured with metered data fluctuate across design choices? (RQ1)

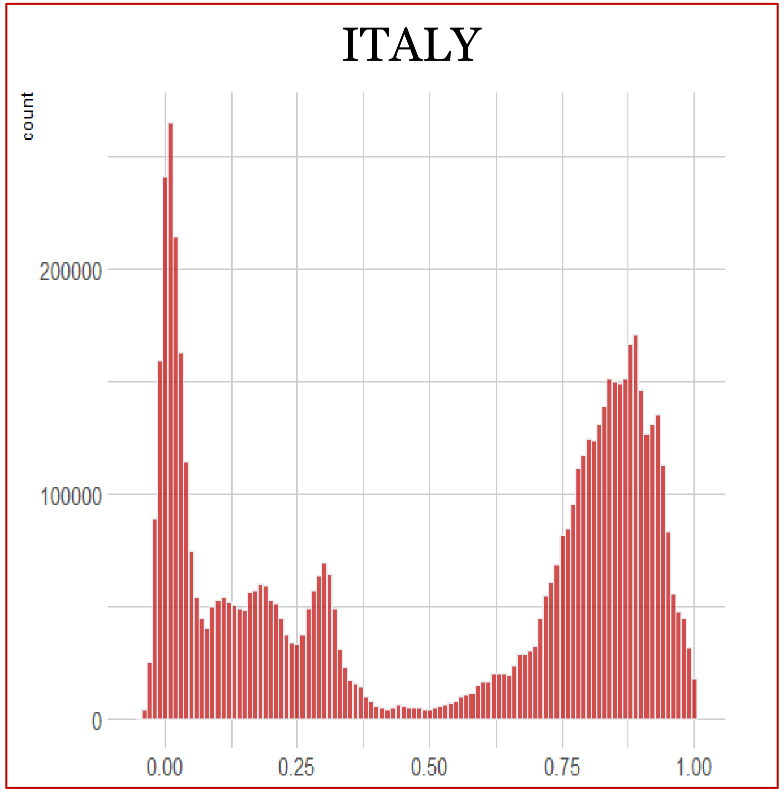
RESULTS

Convergent validity

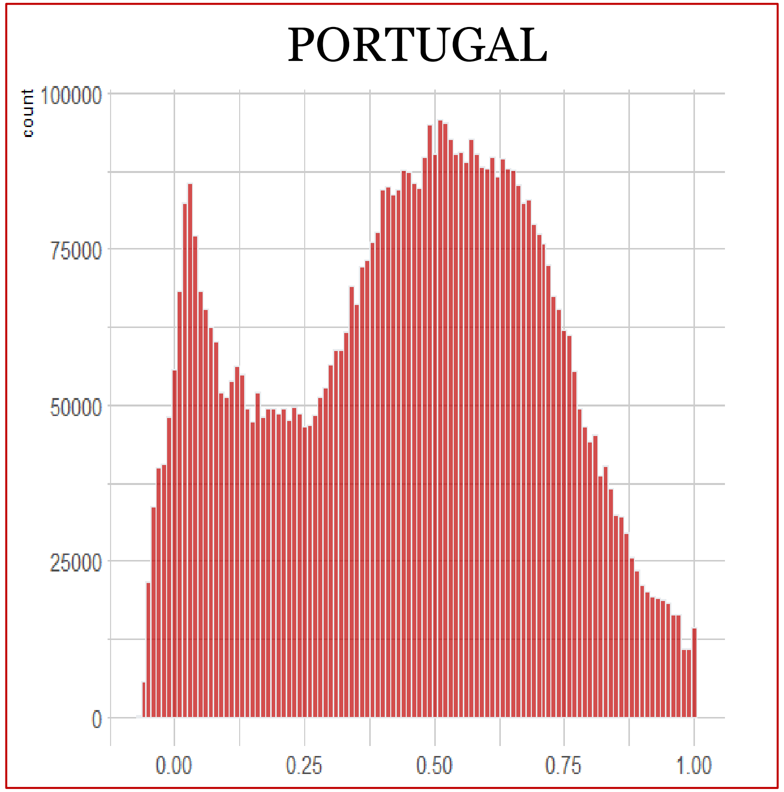
Correlation between different specifications



Mean: .40
Media: .41
1st Quart: .15
3rd Quart: .63



Mean: .51
Media: .69
1st Quart: .10
3rd Quart: .85

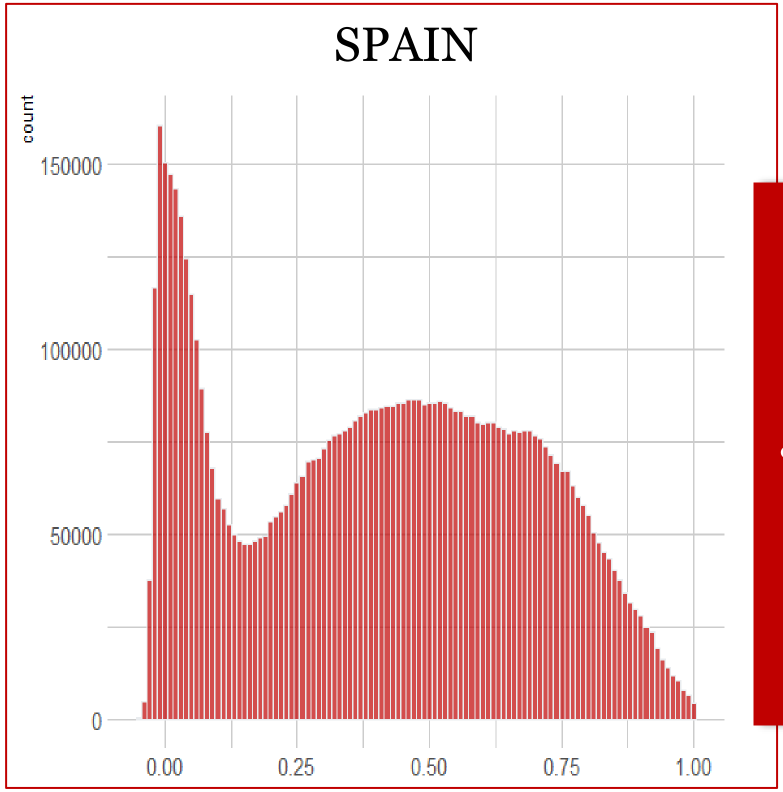


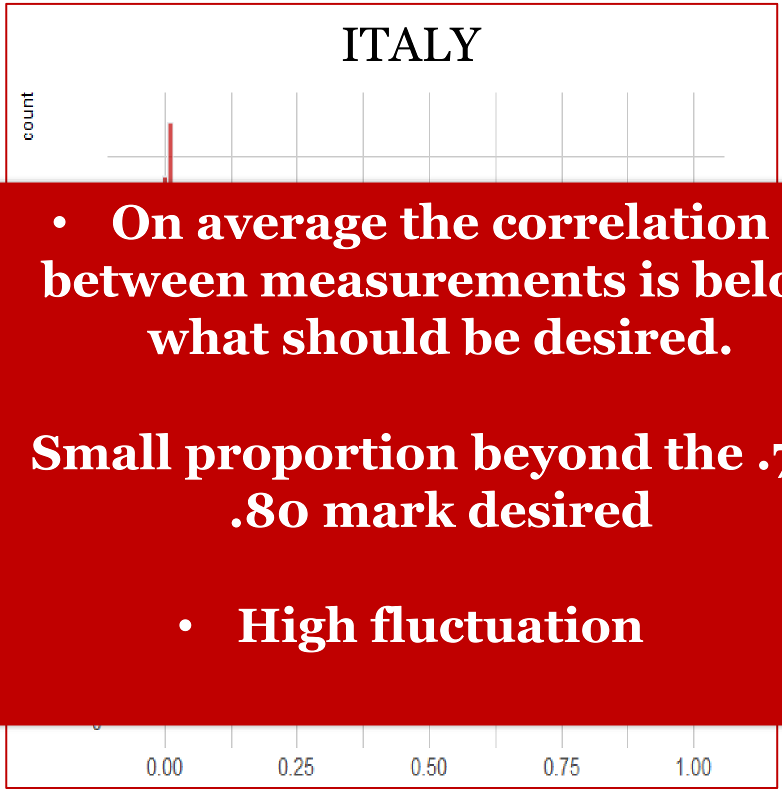
Mean: .45
Media: .47
1st Quart: .24
3rd Quart: .64

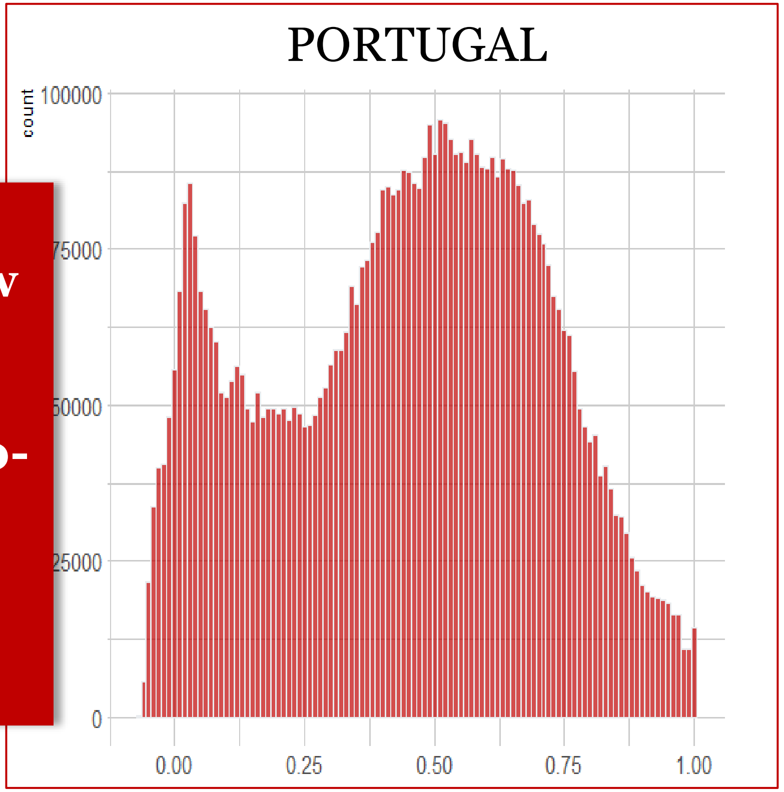
Convergent validity

web data opp

Correlation between different specifications







- **On average the correlation between measurements is below what should be desired.**
- **Small proportion beyond the .70-.80 mark desired**
- **High fluctuation**

Mean: .40
Media: .41
1st Quart: .15
3rd Quart: .63

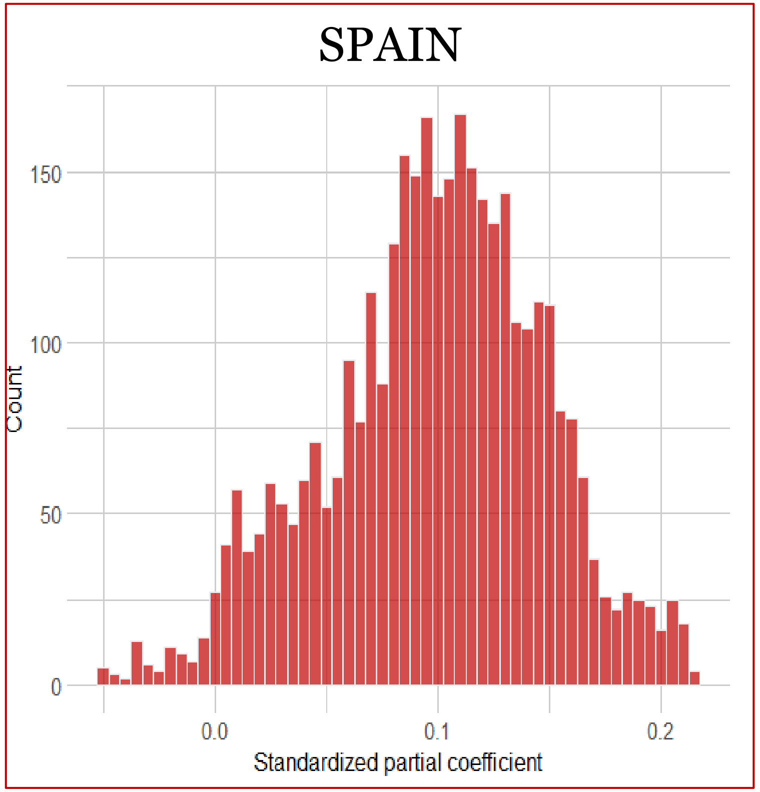
Mean: .51
Media: .69
1st Quart: .10
3rd Quart: .85

Mean: .45
Media: .47
1st Quart: .24
3rd Quart: .64

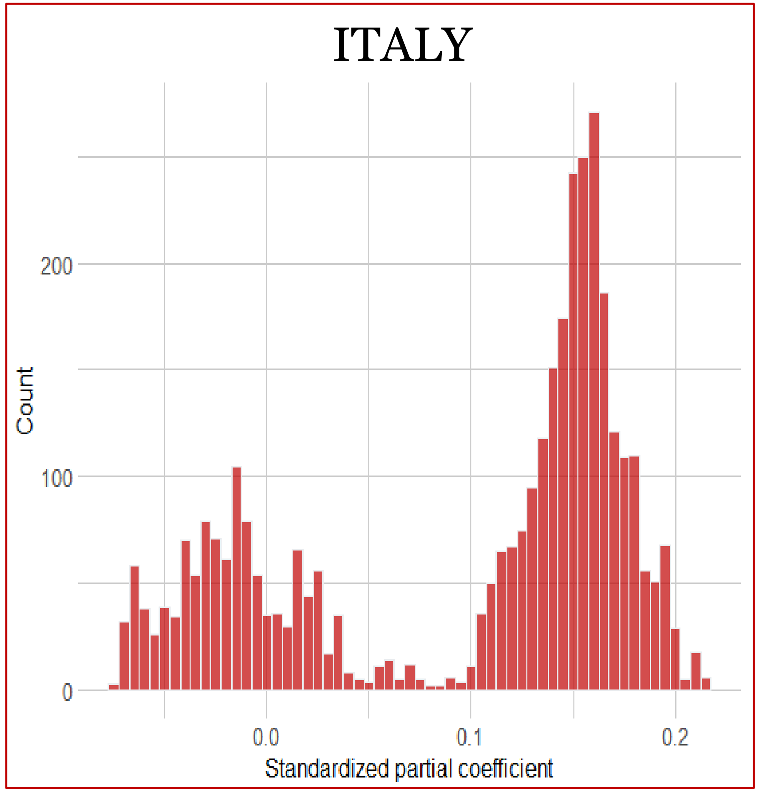
RESULTS

Predictive validity

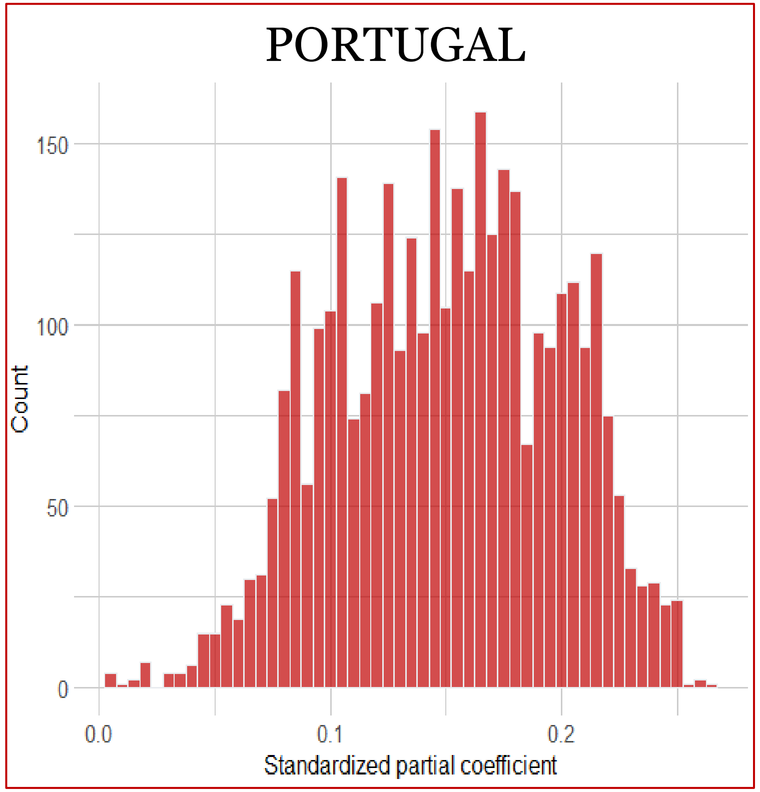
Association with political knowledge across different specifications



Mean: .099
Media: .102
1st Quart: .069
3rd Quart: .132



Mean: .098
Media: .140
1st Quart: .098
3rd Quart: .160



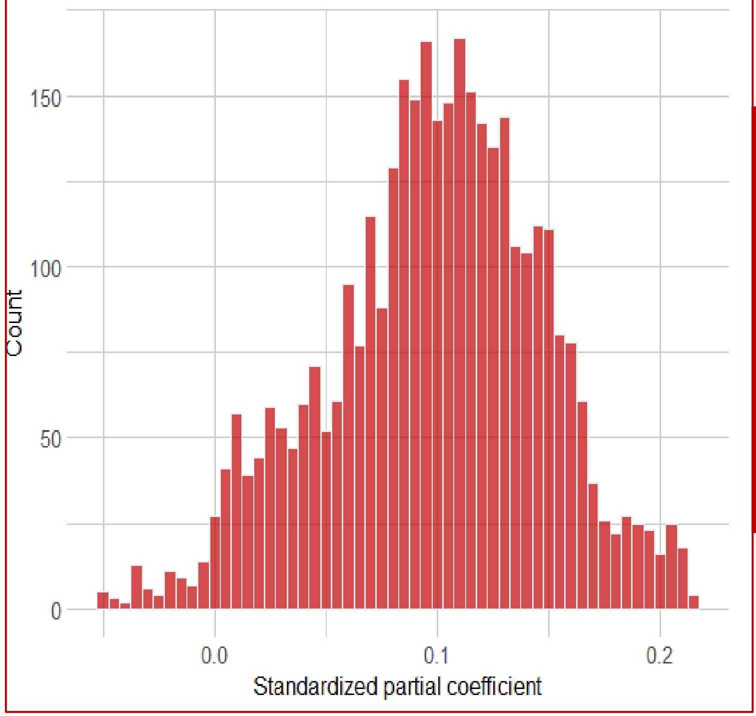
Mean: .150
Media: .152
1st Quart: .113
3rd Quart: .188

RESULTS

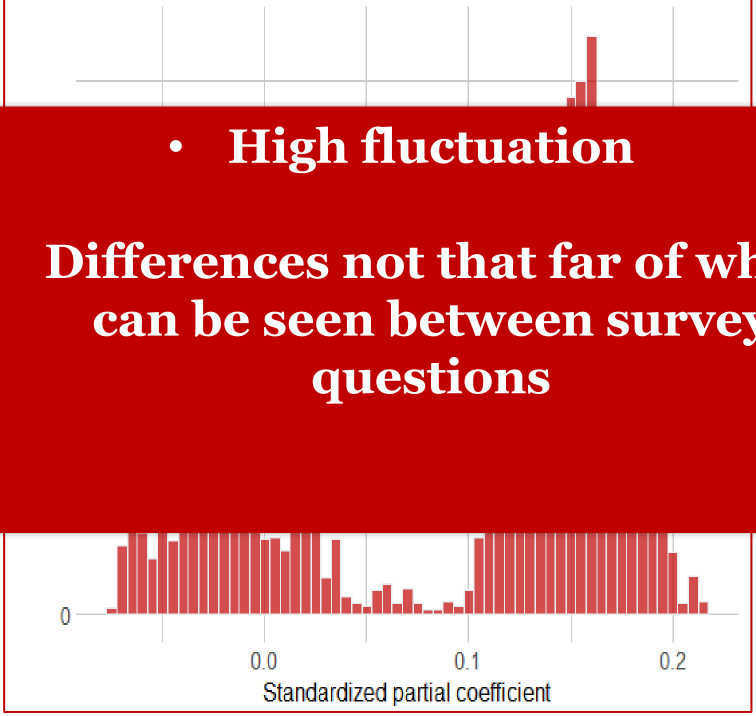
Predictive validity

Association with political knowledge across different specifications

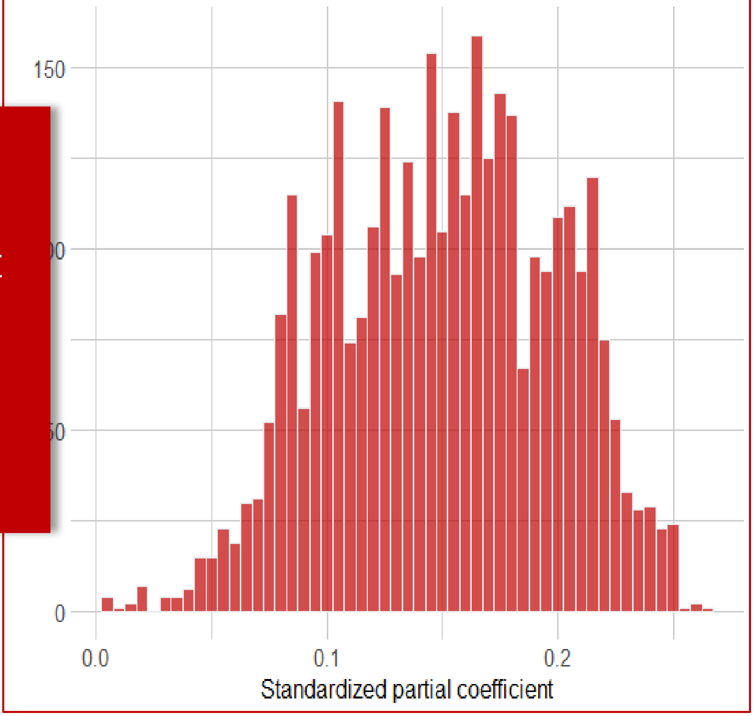
SPAIN



ITALY



PORTUGAL



- **High fluctuation**
- **Differences not that far of what can be seen between survey questions**

Mean: .099
Media: .102
1st Quart: .069
3rd Quart: .132

Mean: .098
Media: .140
1st Quart: .098
3rd Quart: .160

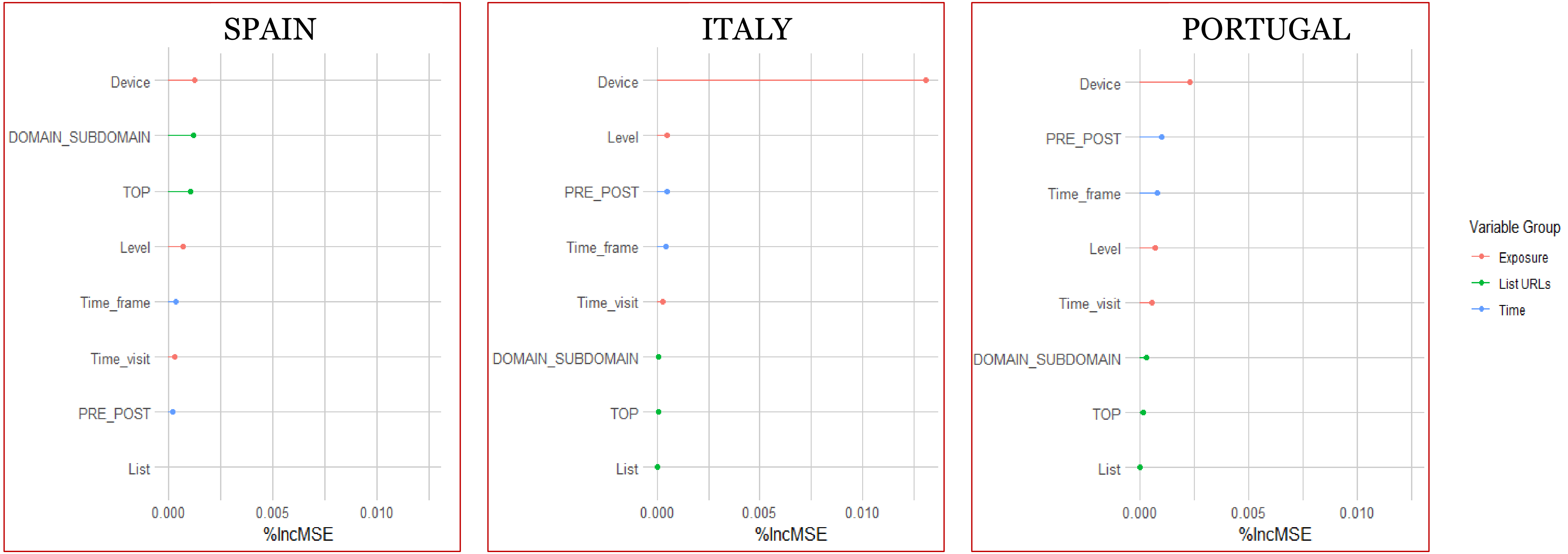
Mean: .150
Media: .152
1st Quart: .113
3rd Quart: .188

What design choices have a higher impact on predictive validity? **(RQ2.1)**

To what extent do different design choices affect the predictive validity of metered data measures? **(RQ2.2)**

RESULTS

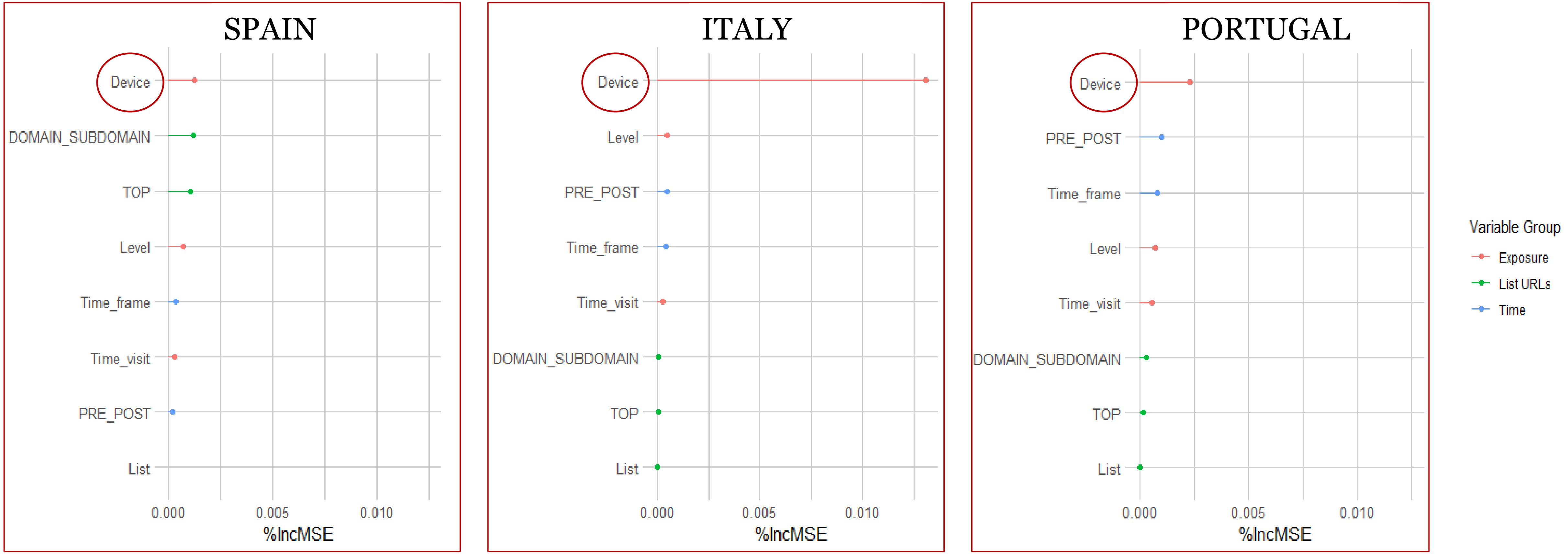
The importance of each design choice



* These results agree with the conditional (unbiased) important measures from cforest

RESULTS

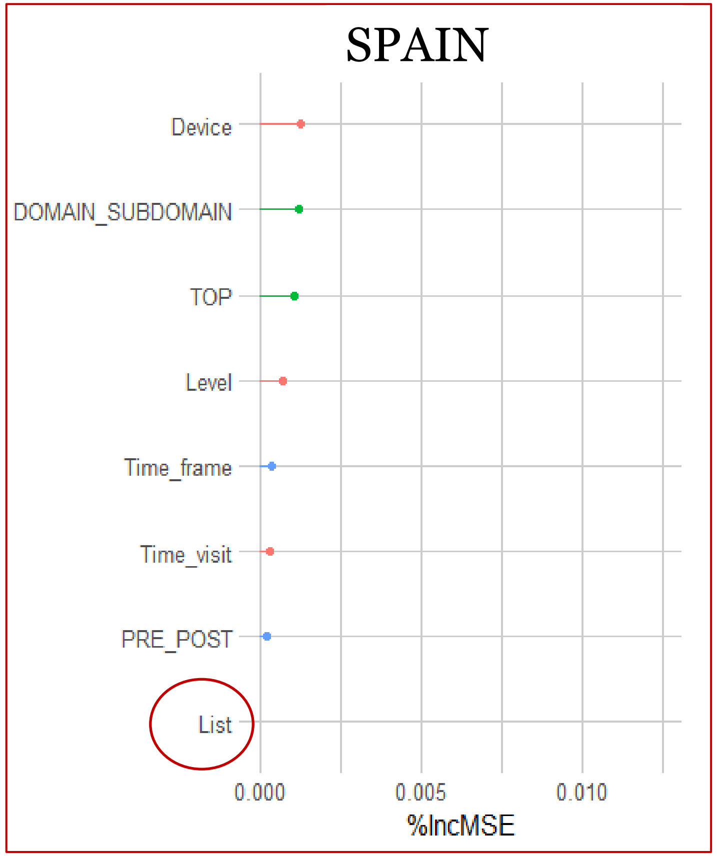
The importance of each design choice

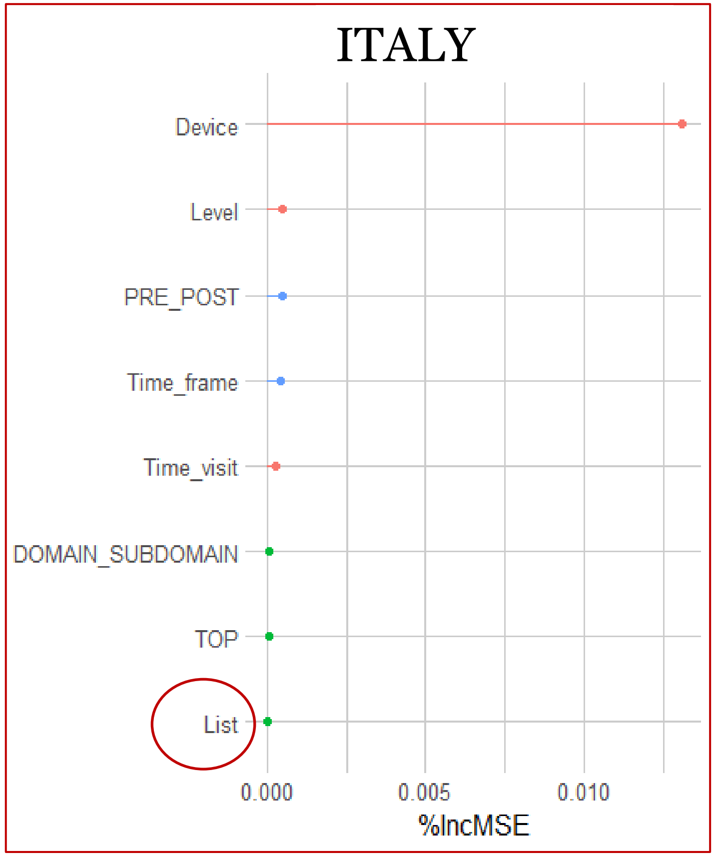


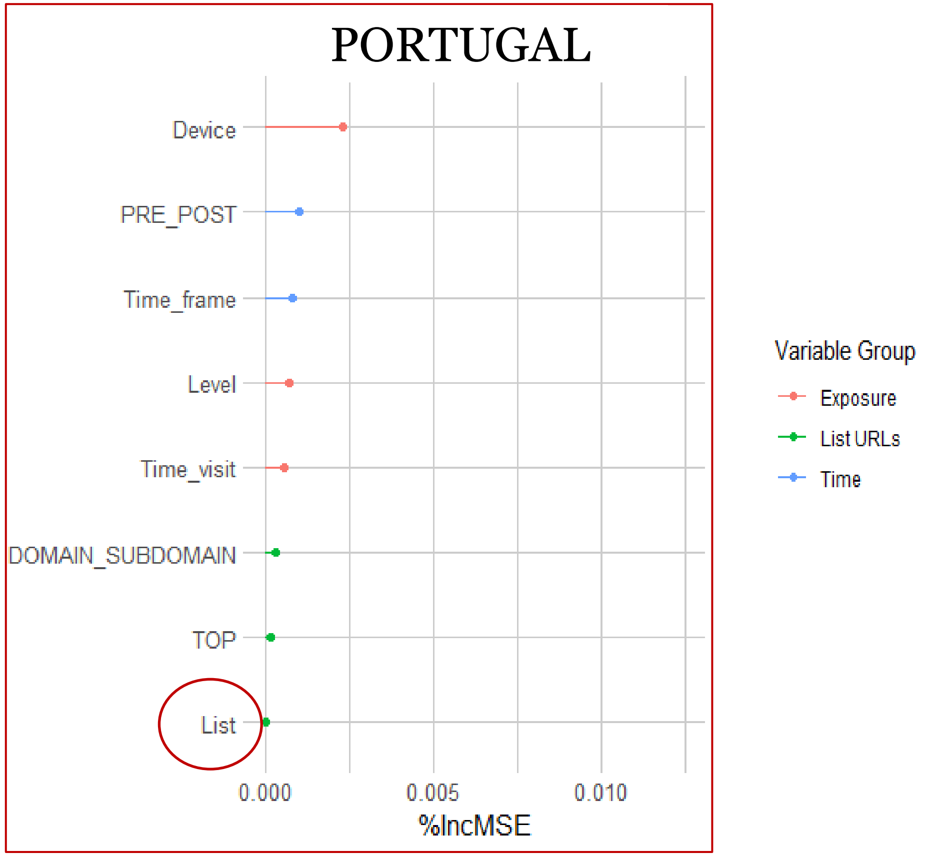
The **device** information used is the **most important variable** across countries

RESULTS

The importance of each design choice



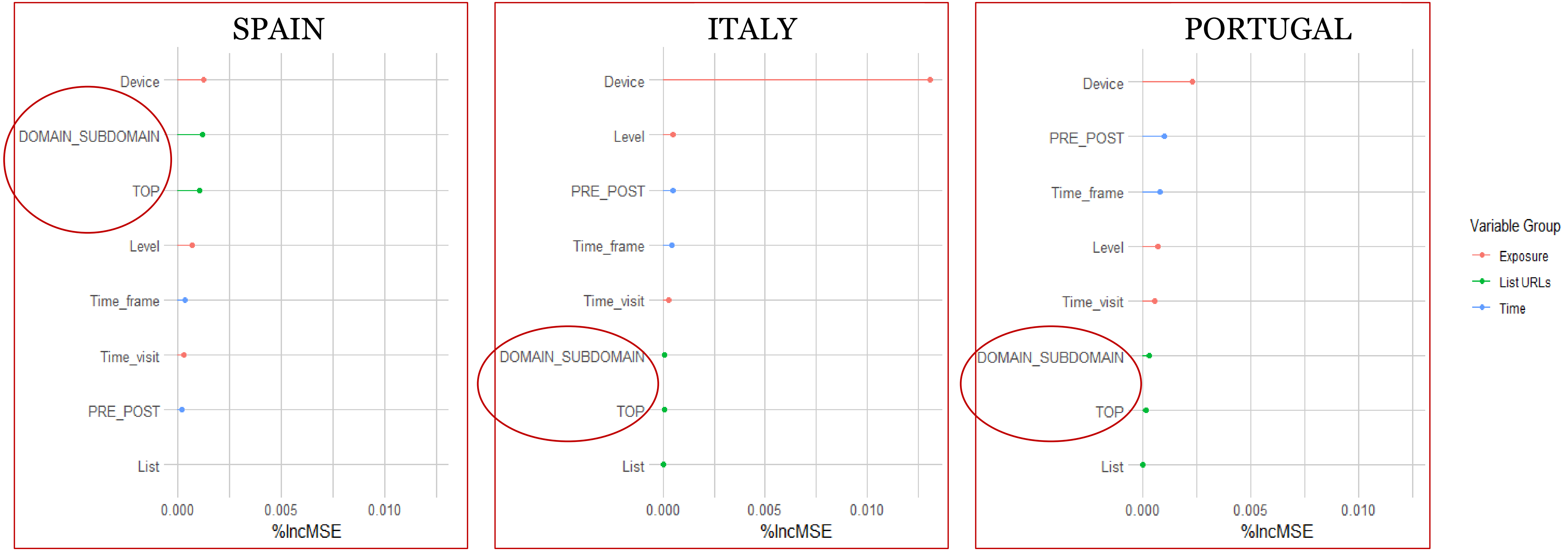




The **ranking list** used is the **less important variable** across countries

RESULTS

The importance of each design choice

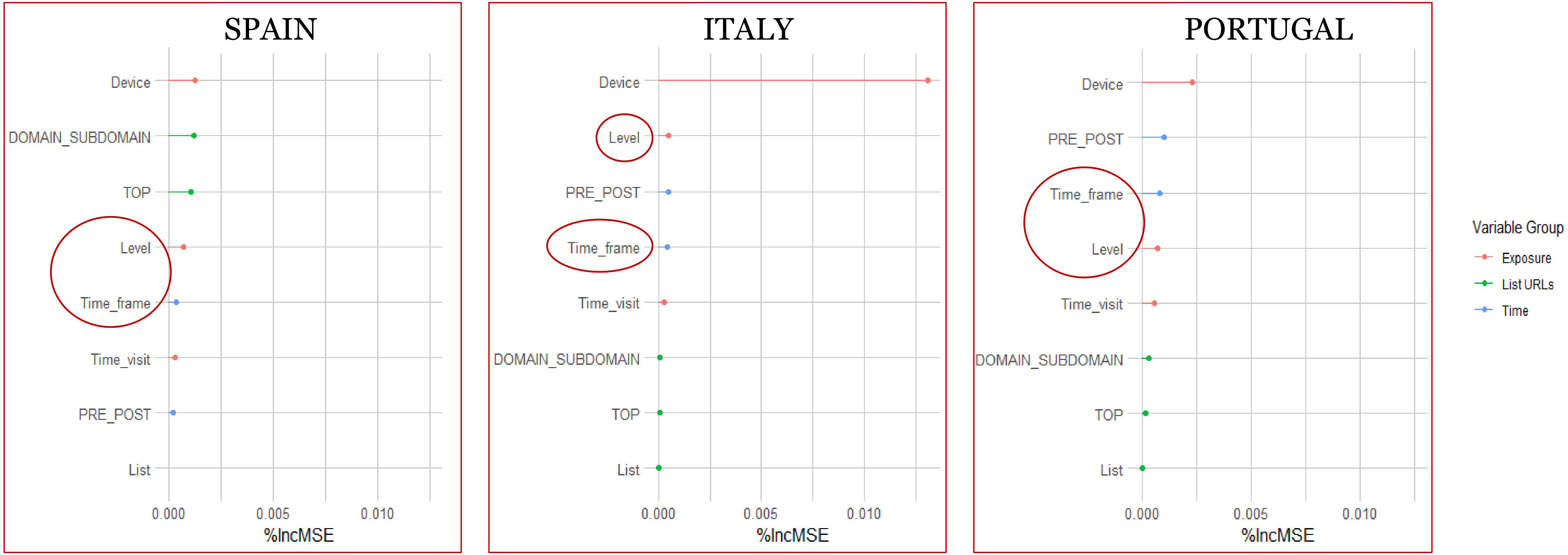


Spain has specific **characteristic** that could explain its differential importance

- More **richness** in the **subdomain** information
- **Regional outlets** (more) **important** in their own regions

RESULTS

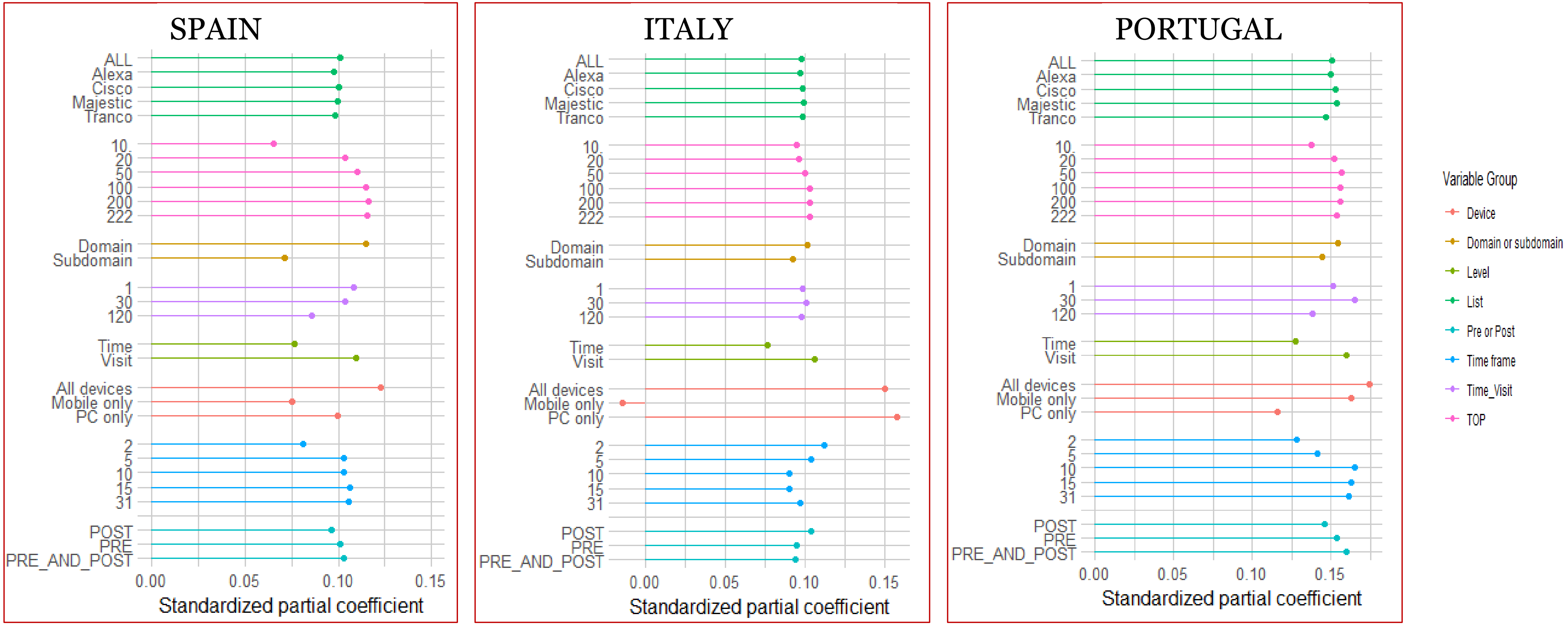
The importance of each design choice



The **level** (visit or time) and the **time frame** (number of days tracked) seem to be **somewhat relevant** across all countries

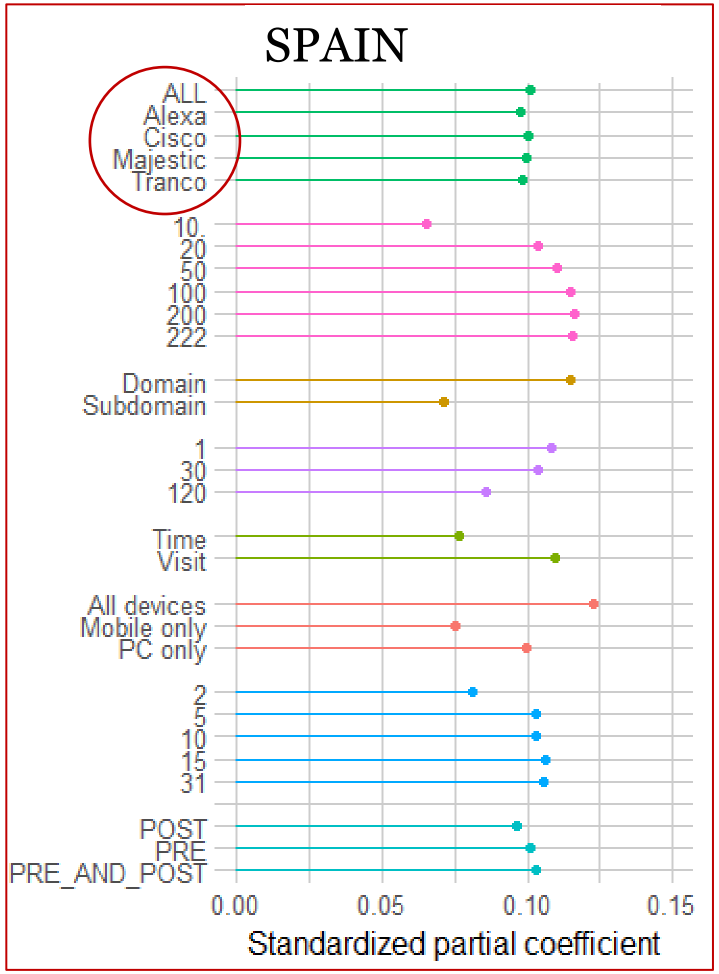
RESULTS

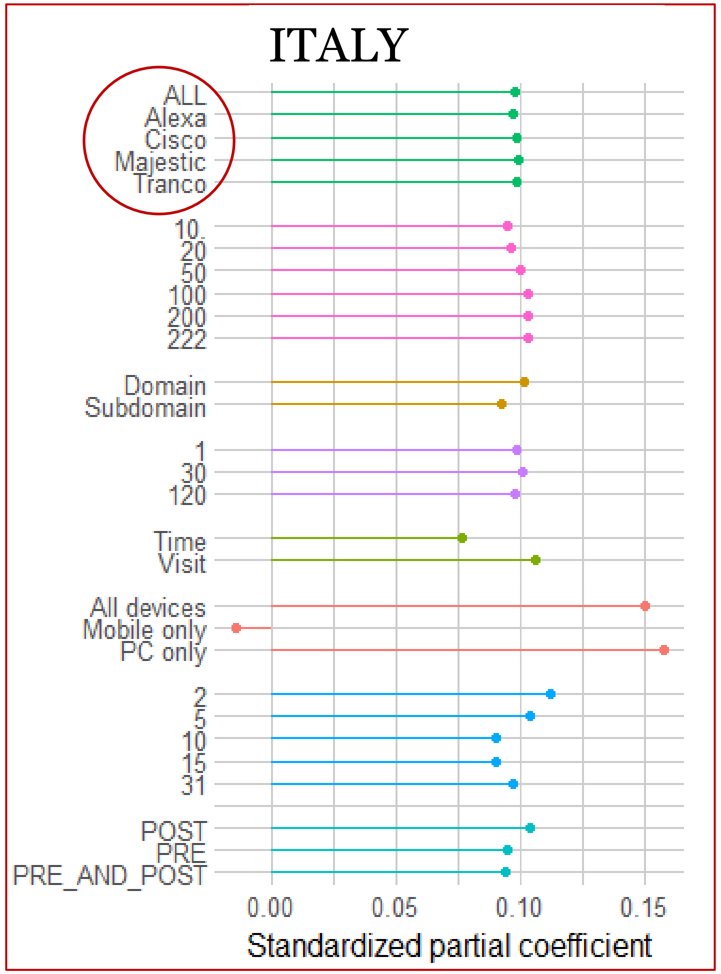
Marginal effect of each specification

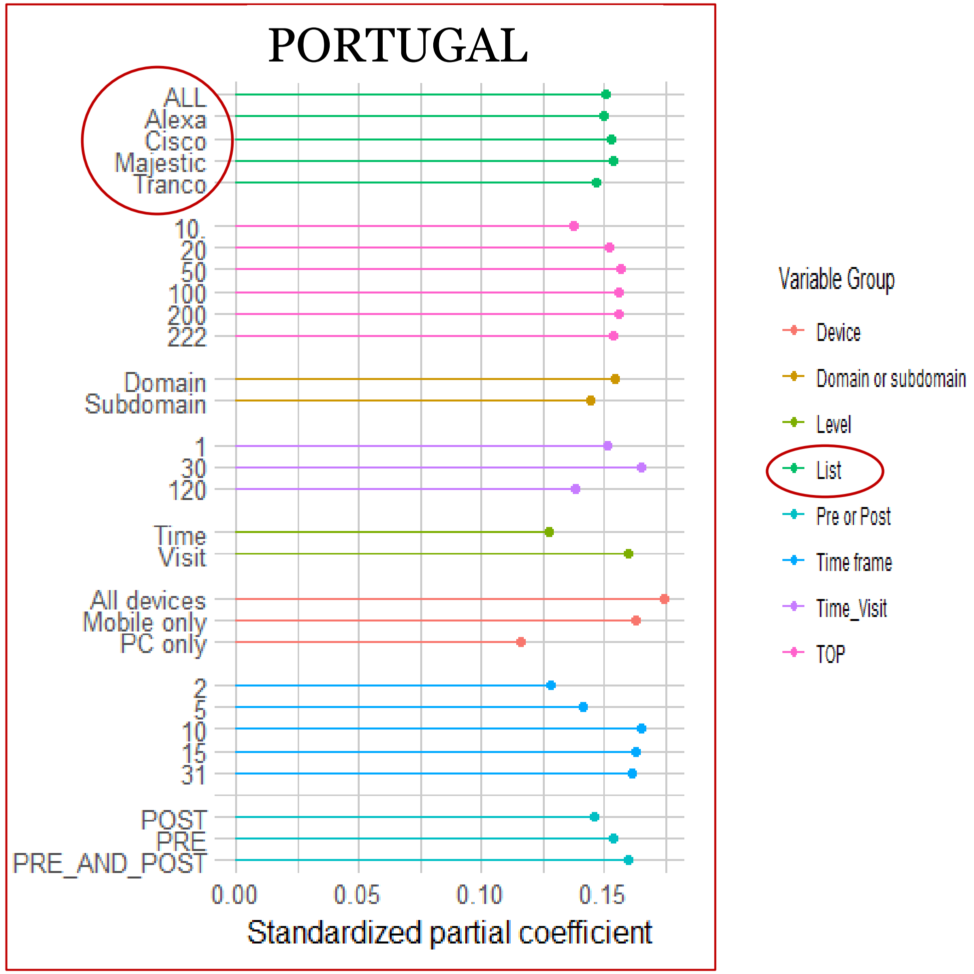


RESULTS

Marginal effect of each specification



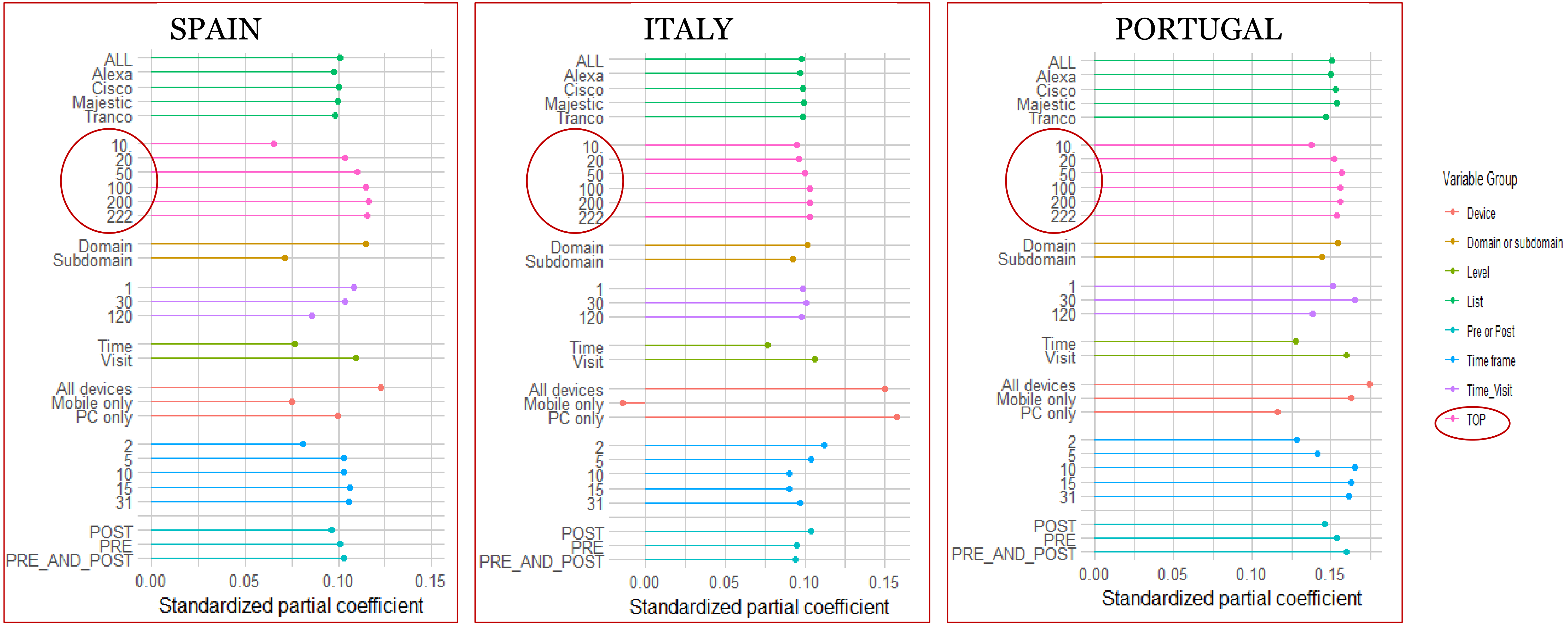




Almost **no fluctuation** across the different **ranking lists**

RESULTS

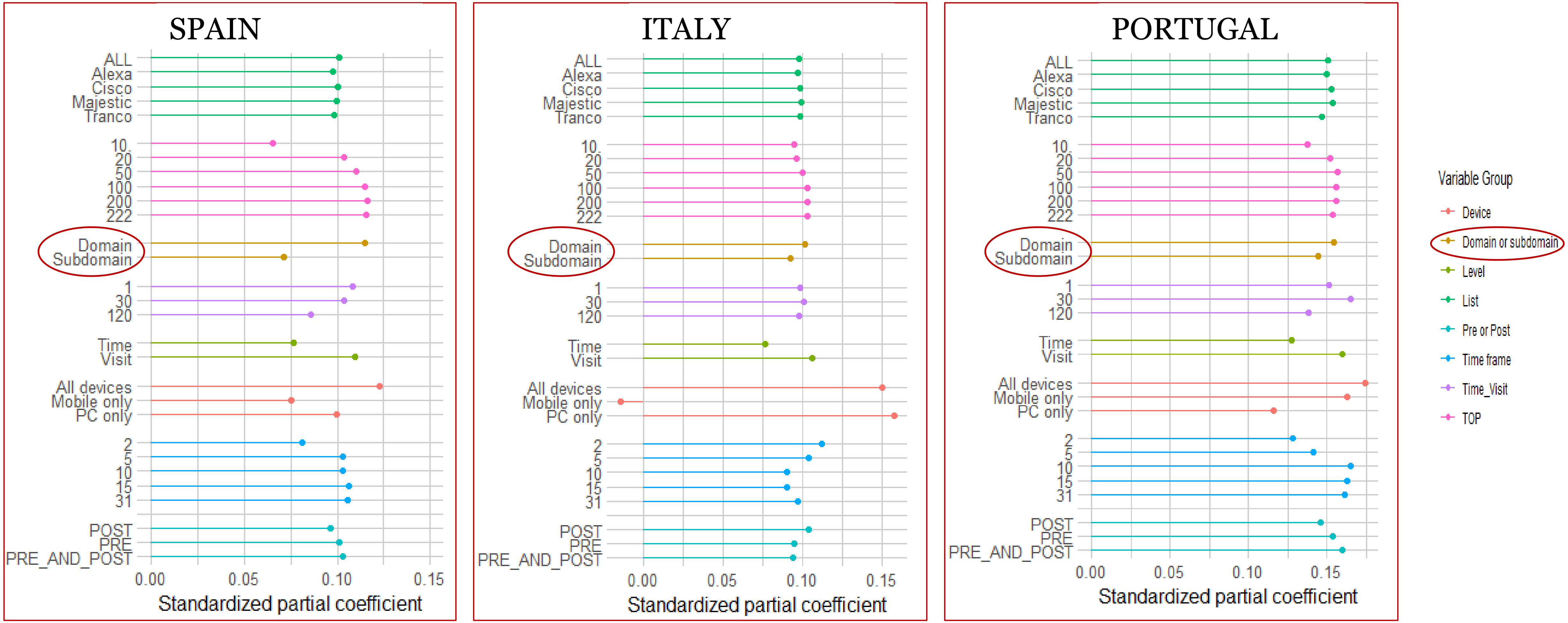
Marginal effect of each specification



Apart from Spain, little fluctuation. The top 50-100 outlets seem to work fine

RESULTS

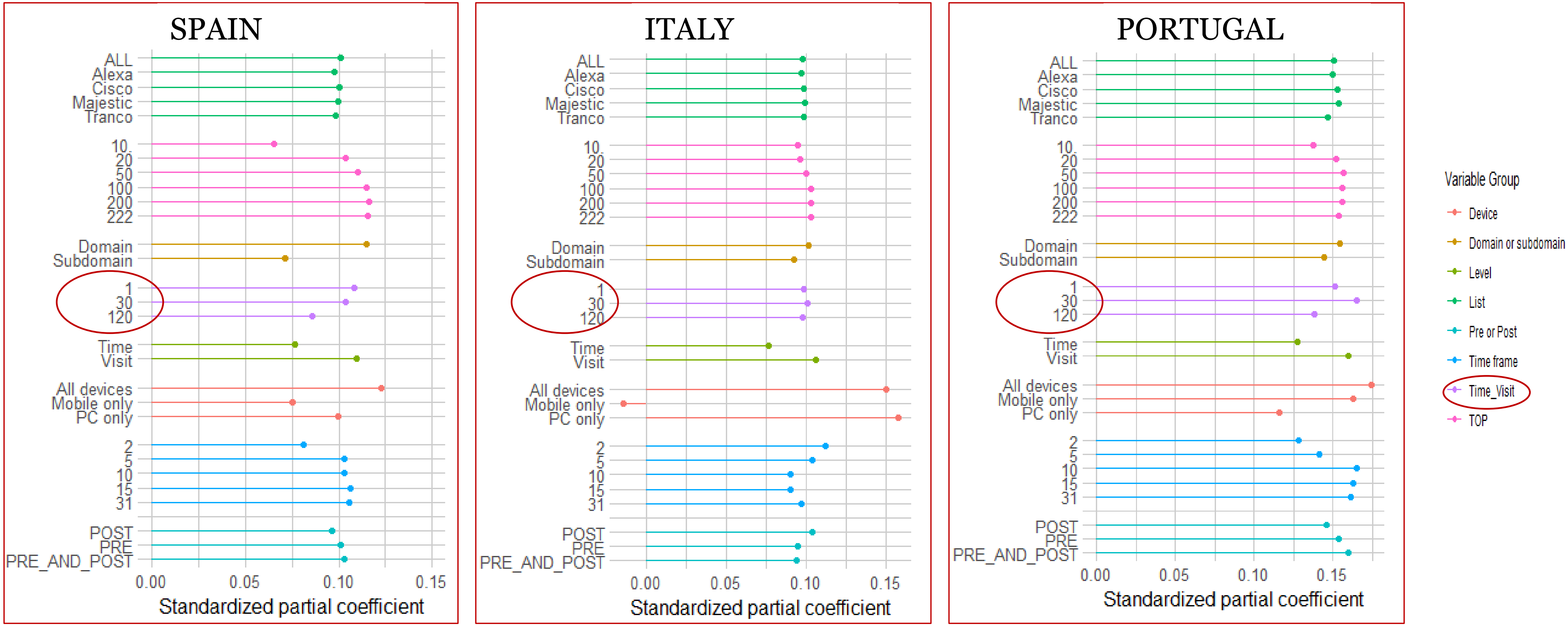
Marginal effect of each specification



Using domain information yields higher predictive power

RESULTS

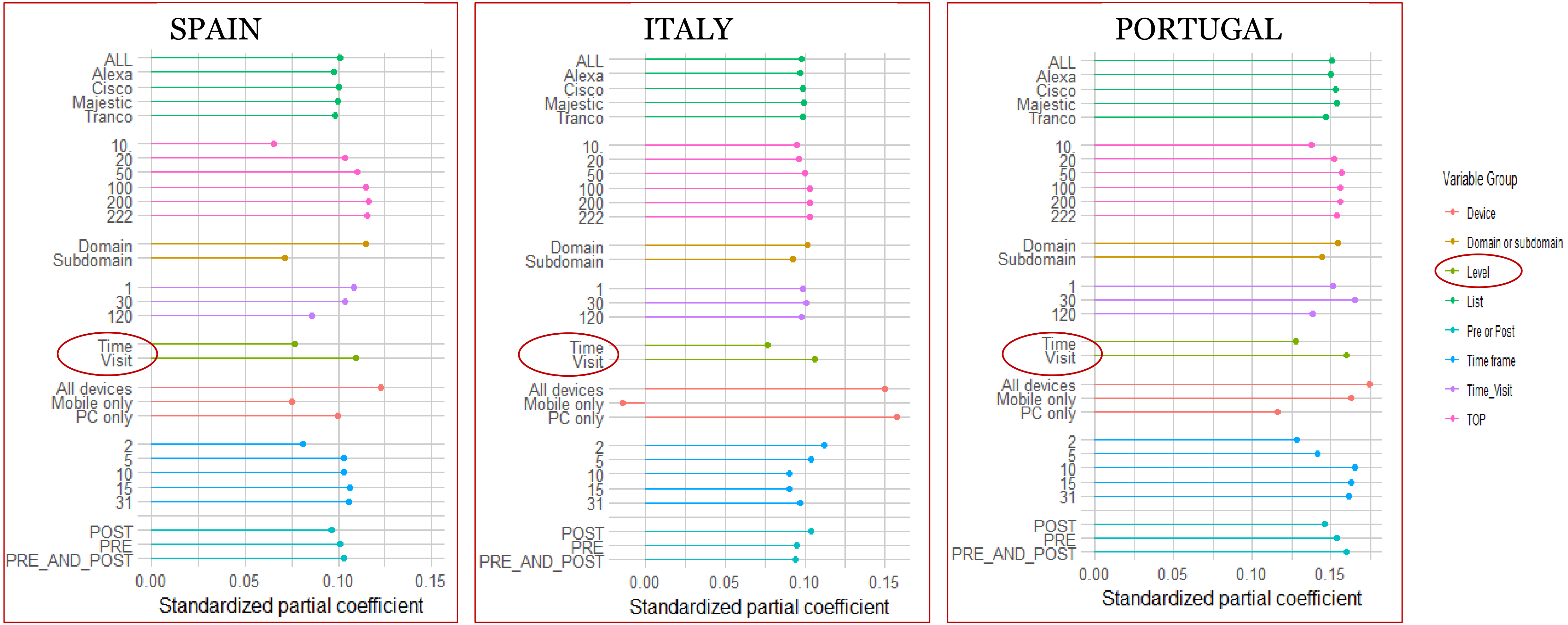
Marginal effect of each specification



Slight but inconsistent fluctuation. The **120s always the lowest**

RESULTS

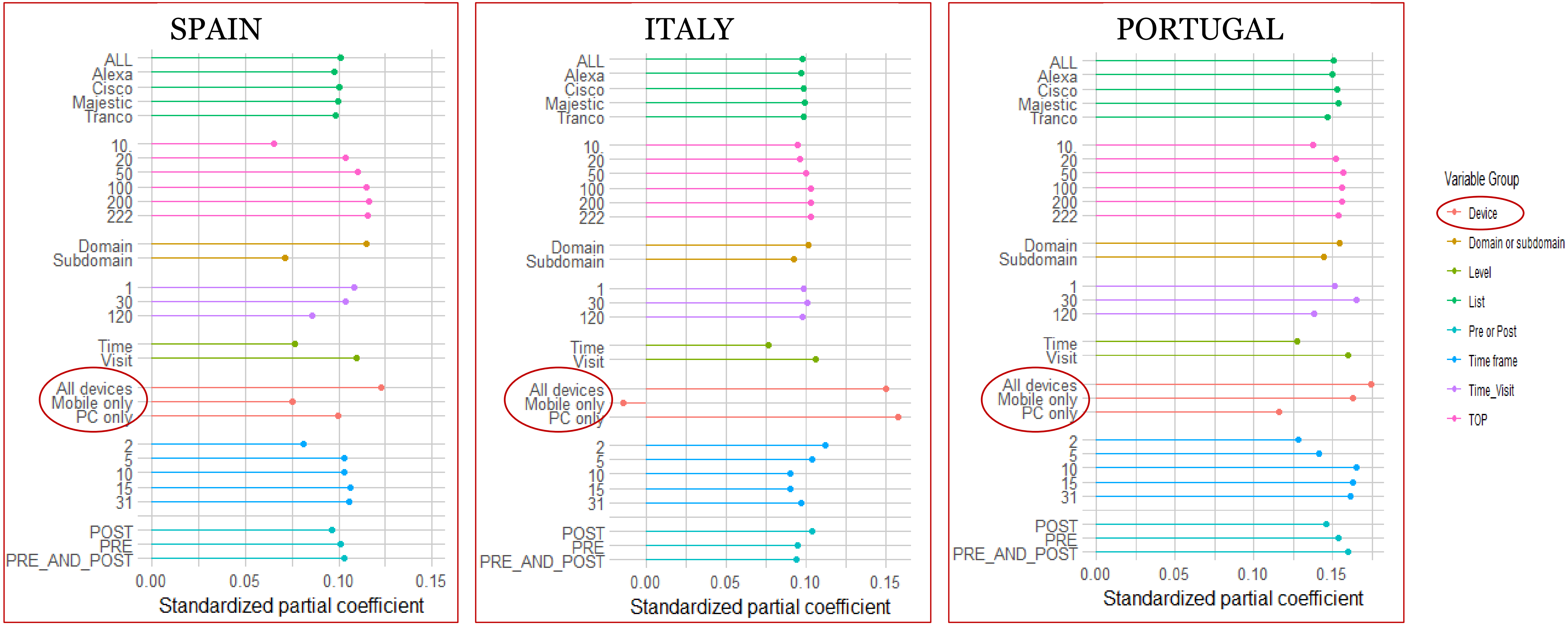
Marginal effect of each specification



Counting visits always leads to higher predictive power

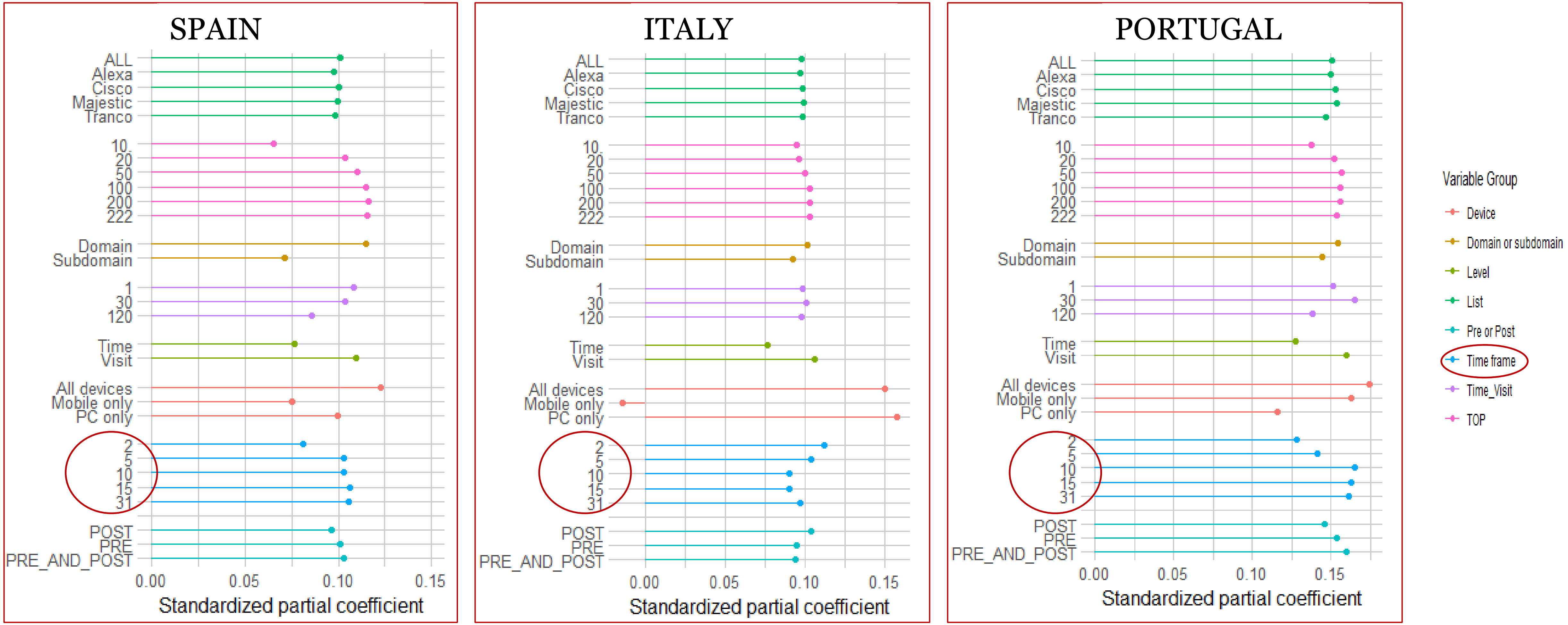
RESULTS

Marginal effect of each specification



Although inconsistent across countries, using information from **both devices seems as the most stable option**

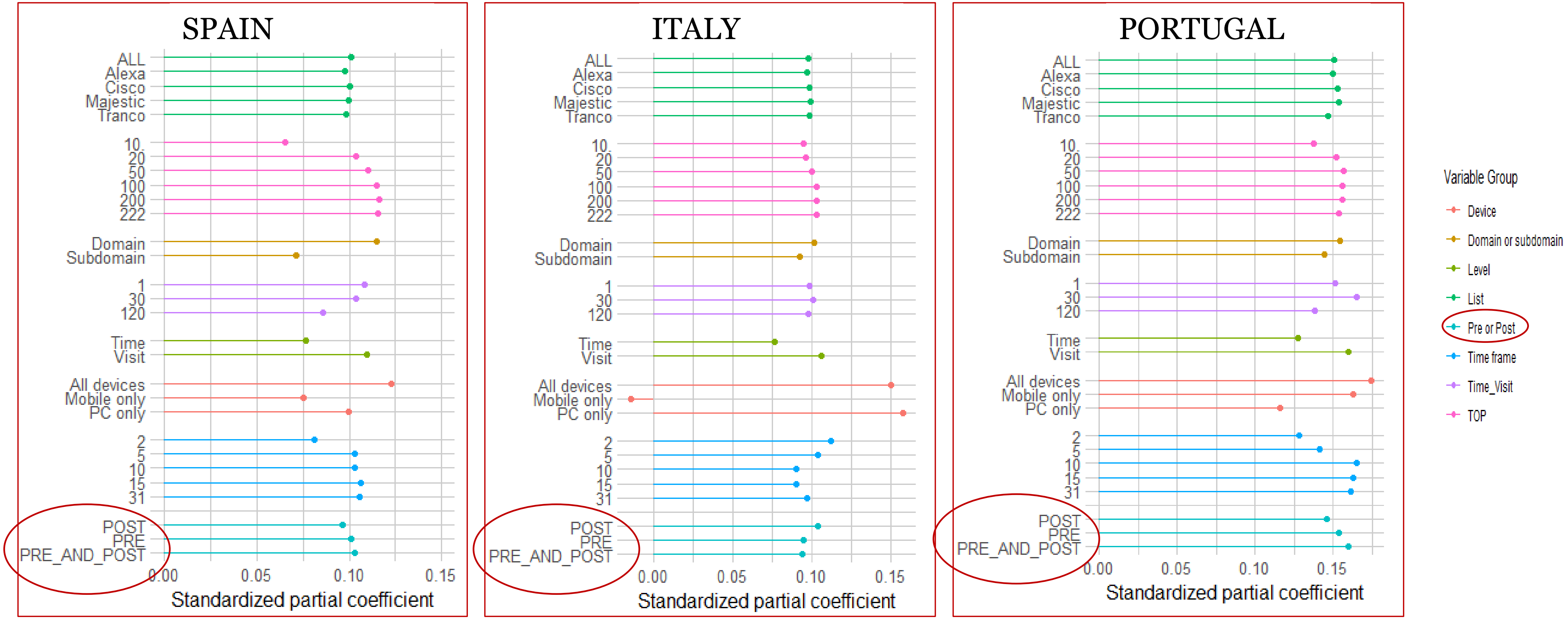
Marginal effect of each specification



The coefficients fluctuate across tracking periods. Italy behaves differently.
10 to 15 days seems to yield the highest predictive power.

RESULTS

Marginal effect of each specification



Little fluctuation. Apart from Italy, using information from after the survey seems to yield lower predictive power.

CONCLUSIONS

Take-home messages

- Many different design choices need to be made when measuring online news media exposure with metered data
- The average-to-low convergent validity + the fluctuation of predictive validity asks for more research…like with surveys!
- Some practical tips
    - Making inferences using only PCs and Mobile devices should be avoided
    - Using the 50 most visited news media outlets from any of the most common ranking lists should work fine.
    - 10 to 15 days of tracking before the survey seems to be a sensible choice

Thanks!

Questions?

Oriol J. Bosch|PhD Candidate, The London School of Economics

o.bosch-jover@lse.ac.uk

orioljbosch

https://orioljbosch.com/

Predictive validity

Measurements generating the highest associations

- **Spain:** Pre | 15 days | PC & Mobile | Visit | 1 second | All news outlets
- **Italy:** Pre | 2 days | PC | Visit | 30 seconds | Top 50 | Cisco
- **Portugal:** Post | 10 days | Mobile | Time | 1 second | Top 50 | Tranco